DOCUMENT RESUME

ED 177 195 TM 009 688

AUTHOR Merwin, Jack C.
TITLE An Author's Perspective of Standardized Achievement Tests.
PUB DATE [Apr 79]
NOTE 14p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, California, April 8-12, 1979)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Achievement Tests; Elementary Secondary Education; *Evaluation Needs; *Measurement Goals; *Psychometrists; *Standardized Tests; *Test Construction; Testing Problems
IDENTIFIERS Teacher Developed Tests; Testing Industry

ABSTRACT

The term, standardized achievement test, has evolved to imply a broad package of materials which includes a wide array of norms, lists of scoring services and educational objectives tested, and aids to interpretation. Test authors are convinced that that individual differences in achievement do exist; that recognition of these differences will lead to improved decision-making about educational experiences; and that such decisions are most helpful when information about achievement level is considered. Four aspects of creating test packages are: (1) which information needs should be addressed by the test--instruction, guidance, etc.; (2) considerations in the development of tests to meet such needs--subject content of the test, type of interpretation of scores, and whether the test is teacher-developed or standardized; (3) difficulties leading to test limitations--testing time, cost effectiveness, and fallibility or poor use of test results; and (4) how the information generated from the test is likely to affect decision making--to corroborate informal conclusions, diagnose educational needs, assess attainment of learning objectives, and to compare students. The standardized achievement test, in spite of its problems, is an important tool which provides unique information about student achievement. (GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Jack C. Merwin

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED177195

IM009 688

An Author's Perspective of Standardized Achievement Tests*

Jack C. Merwin

University of Minnesota

The task I have been asked to do today is to provide a test author's perspective of standardized achievement tests. It is important for me to note that while I am sure there is a general factor on which many test authors would load heavily, the N of 1 sampling in this instance can in no way justify a generalization to a "domain" of authors.

This is an important point. The tests that have been produced make it obvious that there are differences in perspective across authors and that perspectives of authors change over time. Both are important in determining what is available as standardized achievement tests. In my judgement it is desirable that differences across authors does lead to alternatives for potential users. And, it is important that authors attend to both changing information needs and newly developing technology. The changes in tests that have taken place over the years demonstrate that they have. While I am sure that these changes have come about all too slowly for some users and too frequently for others, it is evident that they do occur. A major change has been expansion from a simple package of materials that included the items, directions for administration, a scoring key and a table of norms to a broad package including a wide array of norms, scoring services,

* Prepared for presentation at the 1979 Annual Meeting of the National Council on Measurement in Education

instructional objectives tapped, and aids to interpretation by a variety of educational decision makers. By "standardized achievement test" I will be referring to this broad set of materials.

I believe we authors do share, and continue to base test development efforts over time on the following convictions: 1) that individual differences in the achievements of children and their development do exist; 2) that recognition of these differences will lead students and those who help guide their learning experiences to better educational decisions about course planning, career choice and the selection of educational experiences, and 3) that such decisions will benefit students most when information about levels of achievement is included in the information base used.

I will take the rest of my time then to sketch a perspective of four aspects of the creation of a standardized achievement test package that includes not only the set of questions but all of the aids to administration, scoring, reporting and interpretation that must accompany it. They include: 1) the matter of which of the many information needs the tests should attempt to address; 2) major considerations as authors design and develop instruments for providing such information; 3) some of the major difficulties that must be faced; and 4) a view on how the information generated by use of the test is likely to impact decisions. By doing this I hope I can put a focus on the degree to which an author might expect to accomplish what he sets out to do--i.e., build in strengths--and the limitations on what can be provided through those efforts. First, the matter of information needs to be addressed.

Achievement information needs can be classified in many different ways. Mehrens and Lehmann's classification of instructional, guidance, administrative and research uses of tests is helpful here and can be used to view how information needs have shaped test design over the years. Rice obviously

faced the need for a test that covered common content and could be administered in a somewhat standardized fashion to study achievement in spelling under different instructional modes before the turn of the century. This was followed by the development of achievement tests in specific school subjects. Terman, Kelly and Ruch recognized and responded to a desire for information on individual student variation in achievement across school subjects in the early 1920's. The desire for across group achievement comparisons led to NAEP and has had an impact on the design of interpretative aids as well as tests. While administrative information needs have been with us for many years, the relatively recent recycling of the concerns of reporting to the public on achievement has rekindled attention to these needs. One example of a response to this need is Part IV of the manual for the current Stanford Achievement Tests which is the "Administrator's Guide" to the use of results.

Test development has always followed changes in curricular emphasis and information needs. The various educational constituencies, those who teach and others who use test results take the lead and actually determine what tests will contain. The test author's role is to provide tools that can be used to generate the information these groups see as useful to them. Without such efforts in the past our understanding of individual differences and our ability to identify the gifted and others with special educational needs would not have developed as it has. And, the work must go on as the technology of testing improves and advances in education changes information needs.

Two different types of user needs one cannot avoid in test design and development arise from the desire that the end product be practical to use and as technically sound as possible. Practicality concerns include testing

time and cost as well as ease of administration, scoring and interpretation. And, users rightfully desire tests that are as technically sound as is feasible within limitations set by concerns for practicality.

Test authors face an ever on-going challenge to deal with the inevitable conflicts between practicality needs and technical quality needs. While providing instrumentation to help meet information needs, the author is constantly struggling to find ways to minimize the trade-off necessary--to find ways to meet practical needs better while minimizing sacrifices in technical quality on one hand and to better meet technical requirements without scarificing practical advantages on the other hand.

Let me move then to the second aspect, considerations in the development of tests to meet such needs. Mounting efforts to develop a test that can help the schools meet information needs always involves certain basic considerations including: what information needs are to be addressed (as this determines test content), what type of interpretation of the scores is anticipated, the curriculum content to be covered, and what is to be unique about this test as an information generating device.

One must start with acceptance of the fact that most achievement testing will be done with teacher-constructed tests. Information from these tests, along with that gleaned from observations of the learning process, will be used in making day-to-day decisions regarding the structuring and pace of providing learning experiences. Once, or possibly twice, a year standardized achievement test results are used to confirm and suppliment such information. The subject content is addressed from a broader frame of reference and the tests provide an independent external source of information that can be added to the teacher's information base about the level of achievement of her students.

Fortunately, again, authors of different standardized tests have different approaches that lead to options for the school; some, for example, choosing to include only "basic skills" while others include tests in such subjects as social studies and science as well.

One way to look at design concerns related to anticipated interpretation is in terms of a 2 by N matrix. One dimension is a simple dichotomy of the object of the measurement, individual students or groups of students. On the other axis are types of interpretation including 1) content-referenced (possibly in terms of instructional objectives), 2) self-referenced to examine either change in achievement or realtive strengths and weaknesses at a given point in time, and 3) norm or "other"-referenced. Test design then is guided by the cell or cells of the matrix to be addressed. If only the information of a single cell is addressed, say the cell that calls for norm-referenced interpretation of the scores of individuals, a relatively simple design for questions and interpretative materials is needed. If several cells, say group content-referenced, group norm-referenced and intraindividual self-referenced cells are to be addressed also, a much more complex design is required.

One of the major concerns of authors of standardized achievement tests is the subject matter content of the test. Decisions regarding the curriculum to be covered involves consideration of a complex of factors. The anticipated interpretation and information needs already discussed obviously enter into the decision. Another factor is identification of the curriculum common to a sizable number of schools across the country. That such exists and that it is a large part of any one school's curriculum is evidenced in many ways. One, of course, is the extensive overlap of different sets of curricular materials and scope and sequence charts for major text series.

There is also evidence in the fact that school systems have for years found curricular validity in standardized achievement tests. The position is further supported by noting the success of the consensus approach to objectives development of the National Assessment of Educational Progress that involves teachers, scholars and laymen.

An individual school may well teach for some objectives that are not common enough to be tapped on a national standardized test. And, certainly they will want to develop tests to assess the achievement of these objectives. In light of the movement of 20-22% of the families in our country every year, however, it is fortunate for those children who must change schools that a large common core of objectives does exist.

A particularly challenging content question grows out of the time required to develop a standardized test and conduct the empirical studies involved. This requires analysis of changes then taking place in curricula in attempts to identify those that will catch on and continue in the curriculum. It calls for the sorting out of fad from long term trend. At times one even finds a recycling to what had previously been considered important, as some would claim has been the case with fundamental operations in mathematics at the elementary level.

Based on my observation and experiences with instrument development for both National Assessment and the Stanford Achievement Tests, I can only conclude that the blueprinting of coverage for standardized achievement tests and nationally available instruments aimed specifically at providing results for "criterion referenced" interpretation are very similar.

A related question that faces the test author is the breadth of curriculum the test is to cover. Will it be curriculum of a single grade, two, or possibly even more grades. Here the trade-off is one of breadth for depth,

usually with testing time as a fixed limiting factor.

In addressing the question of what is to make standardized achievement testing a unique information generation source starts with the generally recognized differences between standardized tests and teacher-made tests that appear in most basic measurement texts. The list in Mehrens and Lehmann is typical:

Comparisons between Standardized and Teacher-Made Achievement Test

Characteristic	Teacher-Made Achievement Test	Standardized Achievement Tests
Directions for administration and scoring	Usually no uniform directions specified	Specific instructions standardized administration and scoring procedures
Sampling	Both content and sampling are determined by classroom teacher	Content determined by curriculum and subject-matter experts; involves extensive investigation of existing syllabi, textbooks, and programs; sampling of content done systematically
Construction	May be hurried and haphazard; often no test blueprint, item tryouts, item analysis or revision; quality of test may be quite poor	Uses meticulous construction procedures that include constructing objectives and test blueprints, employing item tryouts, item analysis and item revisions
Norms	Only local classroom norms are available	In addition to local norms, standardized tests typically make available national, school district, and school building norms
Purposes and use	Best suited for measuring particular objectives set by teacher and for intraclass comparisons	Best suited for measuring broader curriculum objectives and for interclass, school, and national comparisons

The principal differences can be summarized into two general categories. First there is the process itself that standardizes content, administration

and scoring so as to permit common interpretations for all who use the test. Such standardization provides opportunity for desired comparisons that simply cannot be made without it. I recall vividly the strong position taken on this point by a technical advisory committee of distinguished scholars when the National Assessment was just starting. They insisted that "carry-over" items would be needed to be administered in two or more assessments to assess "Progress"--that different items written and administered at different times, even when written to specifically stated objectives, simply could not be used to make the desired comparisons.

A second generally unique aspect of standardized achievement tests is the adding of an empirical base to the logical base for item development and selection for inclusion on a final form. The items are written to a carefully developed blueprint, scrutinized by subject matter specialists, those particularly sensitive to cultural and sexual stereotyping as well as measurement specialists and editors for logically sound characteristics. In addition, empirical data regarding reading level and other characteristics are assembled, as are the results of the "proof-of-the-pudding" test of administration of the items at least once, if not more than once. While such extensive logical and empirical review is a time consuming and costly process, it does lead to a sound basis for eliminating relatively poor items in favor of the 50 or less percent of the original set that will appear on the final test. One who has never engaged in such a process can have no real appreciation of the flaws that can crop up in a test item that "looks like" a great item and "appears" to measure a given instructional objective very well.

Let me turn briefly then to several areas of difficulties than can lead to limitations--weaknesses if you will--of the test.

Testing time is a basic and overriding constraint. Many schools operate on class periods of 40-55 minutes and find test administrations that take longer than this disruptive. This in turn poses the need to develop a limited and carefully articulated set of items that will tap what school people feel are important characteristics of the fabric of the subject matter being tested. This is critical. The set must be more than a collection of items addressed to highly specific and unrelated objectives. It must be responsive to important common threads of learning that extend over time.

A second constraint is the understandable user desire for a quality test at the lowest possible cost.

A simple fact of life that poses potential difficulties is that data must be collected in an empirically based development of a test. The technical quality of a test and the norms that can be made available to users depends on the input of many people. It simply cannot be achieved without administration of the test as part of the item tryout and norming processes. The technical quality of currently available tests is the result of the willingness of many people to cooperate in the needed administrations during test development. There is no escaping the fact that the technical soundness of tests to be available in the future will be dependent on such cooperation.

Let me turn next to some "limitations" of standardized achievement tests. First, though some might desire it, standardized tests simply cannot make decisions about kids. While they can be used to obtain information that will assist in the process it is people, not tests, that make educational decisions. It should be noted that many criticisms seemingly directed at tests have actually been criticisms of the use of test results by people.

Test results can make a unique and important contribution to the information base available to the educational decision maker. However, they will never provide all of the information that should be used. While the number and variety of interpretative aids available to those who use test results has increased greatly in recent years, there is a never ending challenge to educate people in their use. With the development of such aids, test authors have taken great care to point out that test scores should always be interpreted in light of all relevant information available.

I am sure that there are some who would consider it a limitation that test scores do not clearly and definitively point to the right alternative or the best policy when educational decisions are faced. If this is what people expect from tests, then this is not only a limitation but an illusion. Experience with program evaluations in recent years has established that a single evaluation study, however broad, can seldom, if ever, lead directly to one best course of action that should be taken without regard to other parts of the relevant information base available to the decision maker. No minimally trained counselor will encourage a counselee to consider test results in isolation. To expect a single test or test battery to provide results that alone would provide all information for an educational decision is simply unrealistic.

One might identify as a weakness that standardized test results are fallible. Authors acknowledge that errors will occur. However, this must be put in the perspective of the fact that all estimates of achievement are fallible and will contain errors. One of the strengths of standardized tests is that authors have provided empirically derived estimates of the reliability and standard errors of measurement which help the user make estimates of the magnitude of errors that might occur. This is unique in that little

If any attention is given to estimates of size of errors in nontest assessments of achievement. And, studies that have been conducted show errors in information collected through such procedures as interviews and teacher judgment are likely to be as large or larger than those found in test scores.

Professional educators know the characteristics and limitations of the information and information generating tools available to them. Based on this expertise, they make decisions and assist students and others who also make educational decisions in the use of the total information base available.

The information provided by standardized achievement tests has in the past and can in the future be useful in many ways. In general, survey achievement test results can be used as an independent check to corrorborate conclusions and confirm hypotheses based on other information. Or, they might call into question those conclusions--possibly pointing to the need to gather more information. As an efficient screening tool they can be used to identify those students needing more involved and time consuming diagnostic testing to arrive at the needed information base.

In recent years, test authors and publishers have attended to changes in the information needs of schools. The tests have been useful in responding to what has been an increasing demand for evidence that children do indeed learn. Some school people find interpretation of results directly in terms of instructional objectives useful. Many standardized achievement tests now available make provision for such interpretation and provide interpretation aids to facilitate it. Additional information is provided when the results can be interpreted in terms of comparisons. This calls for control of content, administration, scoring and interpretation so that meaningful comparisons can be made. It has been specifically to this end that stand-

ardized achievement tests have over the years served as a unique information generation tool. It should be noted that while the availability of normative information adds the possibility of interindividual comparisons, it also paves the way for intraindividual comparisons. It was the genius of Terman and Kelly that led them in the early 20's to recognize that profiles based on a common norm group could provide information regarding the relative levels of achievement of an individual student. While changes in the tests themselves and the accompanying interpretative materials have come about in response to changing information needs, this focus of providing a basis for comparative interpretation has remained a unique strength of standardized tests.

In summary, standardized tests do indeed have both limitations and strengths. That they will not make decision is acknowledged. That they will not provide all of the information needed for educational decisions is also acknowledged. That no test, standardized or not, will tap every objective taught for in a given school is obvious. There will be errors of measurement in standardized test scores, as there will be in scores from teacher-constructed tests and nontest estimates of achievement. And, there must be trade-offs between practical aspects of the use of the test and desired content and technical characteristics. While standardized achievement tests have these limitations they must be judged in light of the unique informtion they can provide and the limitations of alternatives that might be considered.

That standardized achievement tests provide an avenue to unique and important information about student achievement is evidenced in the characteristics of their development. The extensive logical and empirical

developmental efforts to achieve curricular validity and technically sound interpretations is unique. The provision of standard errors of measurement to estimate the probably magnitude of errors is unique and important to appropriate interpretation of results. While results can be interpreted in terms of instructional objectives, the standardization of content, administration, scoring and norms permit unique and important views of achievement through interindividual comparisons and through intraindividual comparisons where test batteries are used. It is these characteristics that make the standardized achievement test an important information generating tool for the professional educator; one that provides unique information about student achievement.